DOCUMENT RESUME

ED 069 796 TM 002 282

| | |
|---|---|
| AUTHOR | Lord, Frederic M. |
| TITLE | Power Scores Estimated by Item Characteristic Curves. |
| INSTITUTION | Educational Testing Service, Princeton, N.J. |
| SPONS AGENCY | National Science Foundation, Washington, D.C. |
| REPORT NO | ETC-RB-72-46 |
| PUB DATE | Oct 72 |
| NOTE | 12p. |
| EDRS PRICE | MF-$0.65 HC-$3.29 |
| DESCRIPTORS | Bulletins; *Mathematical Models; *Probability Theory; Research; *Standardized Tests; Statistical Analysis; Tables (Data); Technical Reports; Testing; *Test Results; *Timed Tests |
| IDENTIFIERS | *Item Characteristic Curves; Power Scores |

ABSTRACT

A method for estimating power scores is described. By way of illustration, it is applied to 21 students who were improperly timed on a standard test. Some empirical results are given in support of the estimation procedure. (Author)

ED 069796

RESEARCH BULLETIN

TM 002 282

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

RB-72-46

# POWER SCORES ESTIMATED BY ITEM CHARACTERISTIC CURVES

Frederic M. Lord

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the author. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
October 1972

Power Scores Estimated by Item Characteristic Curves

Frederic M. Lord

Educational Testing Service

## Abstract

A method for estimating power scores is described. By way of illustration, it is applied to 21 students who were improperly timed on a standard test. Some empirical results are given in support of the estimation procedure.

# Power Scores Estimated by Item Characteristic Curves[1]

Frederic M. Lord

Educational Testing Service

## 1. Introduction

The effect of limiting testing time has been investigated in the past by various empirical studies, set up to treat testing time as a variable manipulated by the experimenter. In many practical situations, time is limited and cannot be manipulated, yet it would be desirable to estimate what the test scores would have been if enough time had been allowed for all examinees to finish.

This might be done easily if test questions were all of equal difficulty and of equal discriminating power. In actual testing situations, however, item characteristic curve theory (Lord & Novick, 1968, chapts. 16-20) is required.

The present report concerns a situation, not unparalleled, in which a group of 21 students was tested on a standard verbal aptitude test under a time limit considerably shorter than should have been allowed. This report describes a tryout of a method for estimating the "power" scores that would have been obtained if the students had had enough time to finish.

In addition to the usual assumptions of item characteristic curve theory, the method assumes that the students answer the test questions in order. Also, that the students respond as they would if given unlimited time--if given more time they would not go back and change answers already given. Such assumptions probably hold approximately for most students, but not all. For this reason, the method outlined here may be of theoretical more than of practical interest.

## 2. Method

The item characteristic function represents the probability $P_{ia} \equiv P_i(\theta_a)$ that examinee a will answer test question (item) i correctly, where $\theta_a$ is the "ability" level of examinee a on whatever psychological dimension is measured by the test. If the test score x is the number of right answers, the expected power score for examinee a on $n_a$ questions is given by

$$\mathcal{E}x_a = \sum_{i=1}^{n_a} P_{ia} \quad .$$

Thus, the power score of examinee a on any n questions can be estimated provided his ability $\theta_a$ and the functions $P_i(\theta)$ can themselves be estimated from available data.

In practice, some specified mathematical form depending on three item parameters, descriptive of item i, is assumed for $P_i(\theta)$, $i = 1,2,\ldots,n$. An essential feature of item characteristic curve theory is that the item parameters remain the same, regardless (within reasonable limits) of the group of examinees. Also, the examinee parameters (the $\theta_a$) remain the same, regardless of the test items administered so long as all items are measuring the same psychological dimension.

The item parameters for each of the $n = 90$ verbal aptitude items composing the mistimed test under consideration were estimated from the answer sheets of a convenient group of 994 students, including the 21 mistimed students. The ability parameter of each of the 21 mistimed

students was estimated from his responses, ignoring any unanswered items at the end of the test. The number-right power score of each mistimed examinee was then estimated from his score on the $n_a$ items actually answered and from $\sum_{i=n_a+1}^{n} \hat{P}_i(\hat{\theta}_a)$, a sum of his estimated $P_{ia}$ computed from $\hat{\theta}_a$ (his estimated $\theta$) and from the estimated item parameters.

## 3. Preliminary Empirical Checks

Granted the two assumptions listed at the end of the first section, the procedure described above can be justified empirically, without considering the assumptions of item characteristic curve theory, if we can show three things:

1. Estimates of item parameters from one group of examinees closely approximate estimates of the same item parameters from other groups of examinees.
2. Estimates of ability parameters from part of a test closely approximate estimates obtained from the entire test.
3. The power score of an examinee on a test can be accurately approximated from his $\hat{\theta}_a$, as estimated from the same test.

A very wide variety of empirical checks will have to be carried out before we can with any assurance outline the circumstances under which these three statements hold. Scattered evidence so far is favorable, as suggested below.

1. Lord (1970) shows good agreement between estimates of item characteristic curves obtained from two rather different groups of examinees. (It is of interest that the two sets

of curves were obtained by two different methods, one of which does not involve any of the usual assumptions of item characteristic curve theory.)

2. Ability parameters for 2,857 examinees were estimated from their responses to items 41-90 of a 90-item verbal aptitude test (a parallel form to the test that was mistimed). The same parameters were reestimated from the same answer sheets, this time from items 1-29, 41-75. After replacing all $\hat{\theta} < -3$ by $\hat{\theta} = -3$ (for 39 values of $\hat{\theta}$), the product moment correlation between the two sets of estimates was found to be 0.944. This value should be compared to the correlation of 0.957 from the same answer sheets between number-right score on items 41-90 and number-right score on items 1-29, 41-75. (The correlations considered in this paragraph are high because much of the same data is used to determine both of the variables correlated. This does not invalidate the present line of reasoning, since a similar overlap is involved when we estimate the $\theta$ of a mistimed individual for the whole test from his performance on items to which he responded.)

3. The product moment correlation between $\sum_{i=1}^{n_a} \hat{P}_i(\hat{\theta}_a)$ and number-right score on a 60-item arithmetic reasoning test was found to be 0.981 for 2,946 examinees, where $n_a$ is the last item answered by examinee a . On a 90-item verbal aptitude test (parallel to the mistimed test), the correlation was found to be 0.992 for 2,926 examinees. The scatter plot for each of these shows a highly linear relationship, with the points grouped closely around a straight line going through the origin. The plot for the 90-item verbal test is shown in Figure 1.

These last results support the conclusion that number-right scores for a set of data can be reproduced with high accuracy from item parameters and ability parameters estimated from the same data. Results cited earlier support the conclusion that, when necessary, the parameters for these data can be approximated by estimates obtained from other sets of data. Thus, power scores can be estimated for examinees who do not have time to finish the test.

Under the mathematical model used, the reason individual points in Figure 1 fail to fall along a perfectly straight line is that some examinees are lucky and some are unlucky in answering the particular n questions administered. The model is a probabilistic one, encompassing these chance fluctuations. The fact that parameter estimates have been used for predicting number-right score instead of true values tends to decrease the scatter of points, since the estimates were made to fit the data.

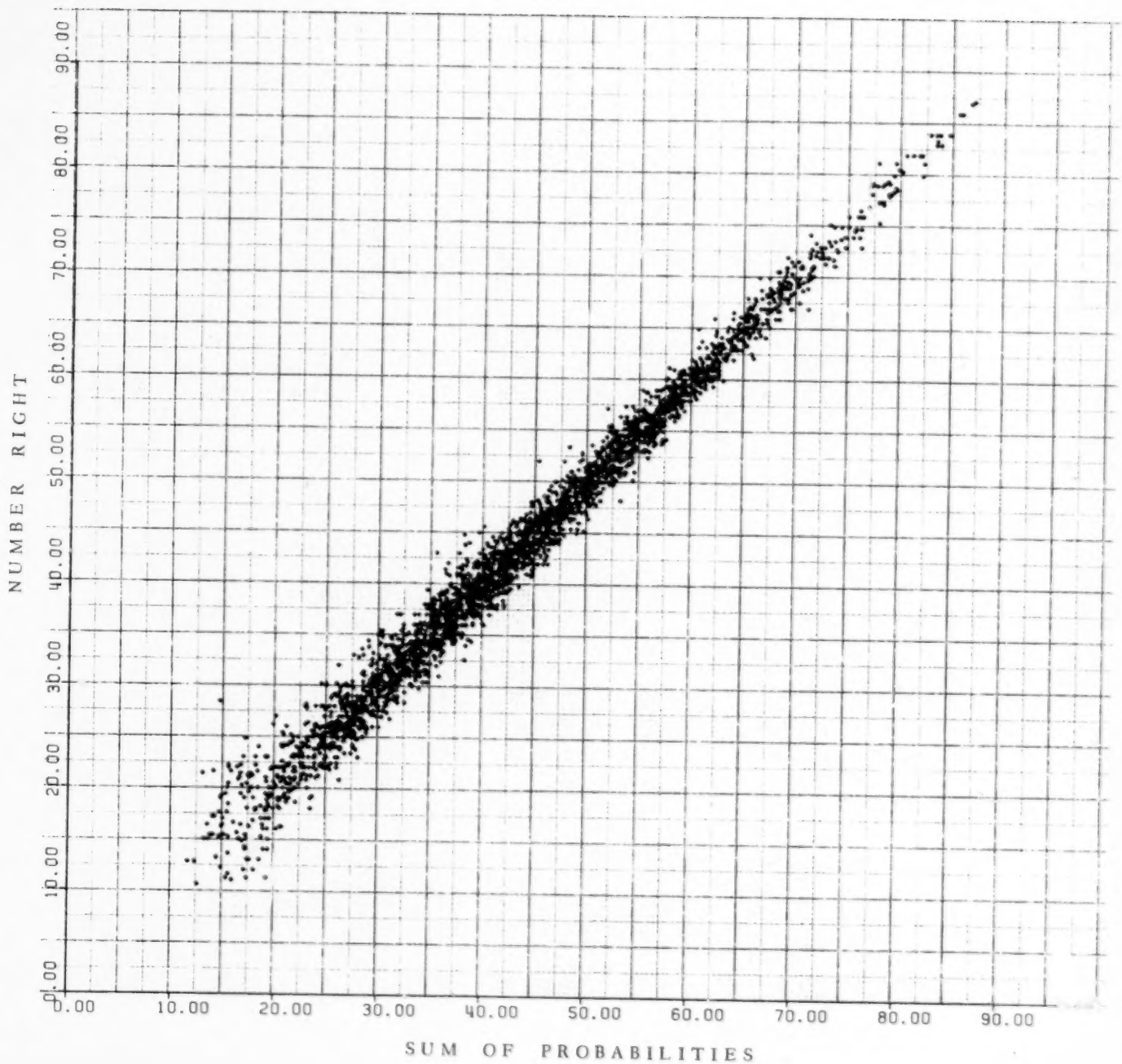


Figure 1. Number-right scores predicted from estimated ability and item parameters.

## 4. Estimated Power Scores

Table 1 shows for each of the 21 mistimed students his number-right score $x_a$, his last item answered $n_a$, his estimated ability level $\hat{\theta}_a$, and his estimated power score

$$x_a + \sum_{i=n_a+1}^{90} \hat{P}_i(\hat{\theta}_a)$$

on the verbal aptitude test. The 21 students are arranged in order of estimated ability.

Since 5 of the students completed the entire test in spite of the shortened testing time, the estimated power scores are of interest only for the remaining 16 students. No empirical check is available for these 16 estimated power scores. The estimates should be valid, however, as long as the students would not use additional testing time to change responses they have already given or to answer questions they have omitted.

The main evidence supporting this last assertion is provided by results such as appear in Figure 1. The predictions for the 16 students should be as accurate as the predictions in Figure 1, except for the fact that the $\hat{\theta}$ for these 16 students were estimated from only $n_a$ of the 90 test items.

Table 1

Estimated Power Scores for Mistimed Students

| Number-right score | Last item answered | Estimated ability | Estimated power score |
|---|---|---|---|
| 68 | 90 | 1.51 | 68 |
| 67 | 90 | 1.51 | 67 |
| 52 | 85 | .94 | 55.6 |
| 47 | 72 | .93 | 59.8 |
| 55 | 90 | .89 | 55 |
| 55 | 90 | .69 | 55 |
| 48 | 81 | .63 | 53.7 |
| 51 | 88 | .60 | 52.4 |
| 45 | 82 | .47 | 49.8 |
| 44 | 83 | .41 | 47.8 |
| 43 | 85 | .20 | 45.7 |
| 39 | 88 | -.03 | 40.2 |
| 37 | 81 | -.07 | 41.4 |
| 38 | 82 | -.16 | 41.8 |
| 39 | 87 | -.39 | 40.2 |
| 32 | 79 | -.41 | 36.9 |
| 32 | 90 | -.42 | 32 |
| 31 | 84 | -.76 | 33.2 |
| 29 | 85 | -.78 | 30.8 |
| 23 | 85 | -1.36 | 24.4 |
| 24 | 87 | -1.46 | 24.7 |

References

Lord, F. M. Item characteristic curves estimated without knowledge of their mathematical form--a confrontation of Birnbaum's logistic model. Psychometrika, 1970, 35, 43-50.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Footnote

[1]Research reported in this paper has been supported by grant GB-32781X from National Science Foundation.